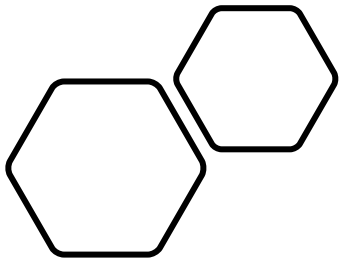
Stats Crash Course

Neil Yetz, M.P.H.

Psy 652

09/30/2020

Objectives



- Learn stats!
- We will learn about:
  - Levels of measurement
  - Descriptive statistics
  - Common stats terminology

Part 1: Variables

Scales of Measurement

- **Categorical variables**: They take on categories. However, none of these categories take an order
  - Gender identity
  - Sex
  - Race/Ethnicity
- **Quantitative variables**: Variables have differing levels. One is less than 2. Two is less than 3, and so on…
  - Income
  - Age
  - Height

Variables can take on FOUR forms

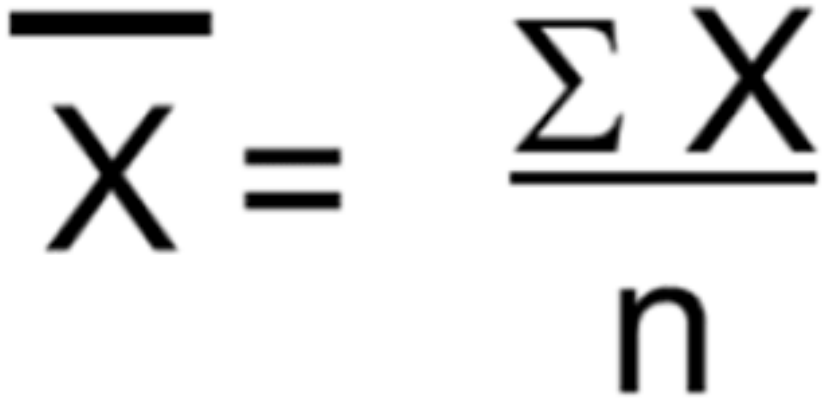
- ***Categorical Variables***
  1. **Nominal:** Completely categorical… But not ordered. (i.e. bird species)

- ***Quantitative Variables***
  2. **Ordinal:** Numbers are ordered… But the order is not equal (i.e. marathon race rankings)
  3. **Interval:** Numbers are ordered AND equal… But there is no true zero point (i.e. Likert scales)
  4. **Ratio:** Numbers are ordered AND equal AND there is a true zero point (i.e. height)

Levels of Measurement

Nominal	Ordinal	Interval	Ratio
"Eye color"	“Race Ranking”	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

Part 2: Descriptive Statistics



Measures of Central tendency

- **Mode:** Most common score
- **Median:** The middlemost score of a distribution of scores
- **Mean:** AKA average, it the sum of the scores divided by the number of scores.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ **Population Variance**

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ **Sample Variance**

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$ **Sample Standard Deviation**

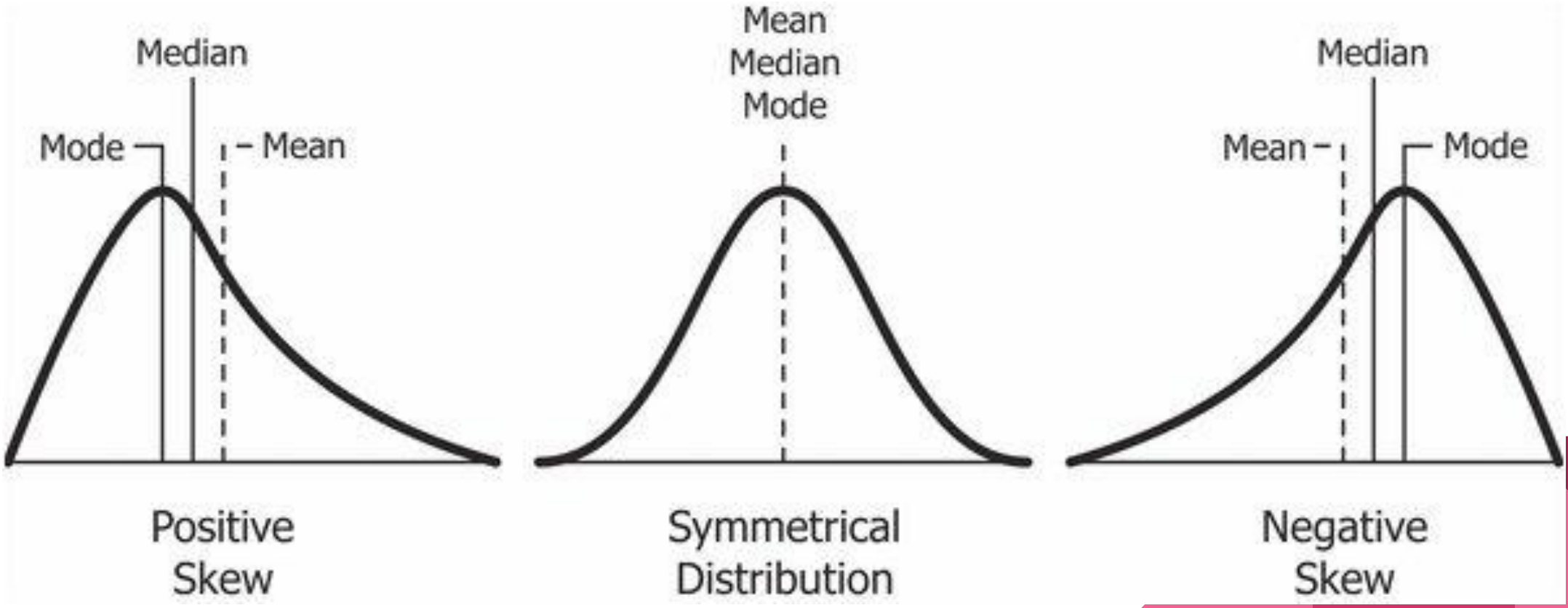
Measures of Variability

- **Variance:** A computation that quantifies how spread out the data is from the mean
  - However, it is in “Squared deviations” meaning it is hard to interpret

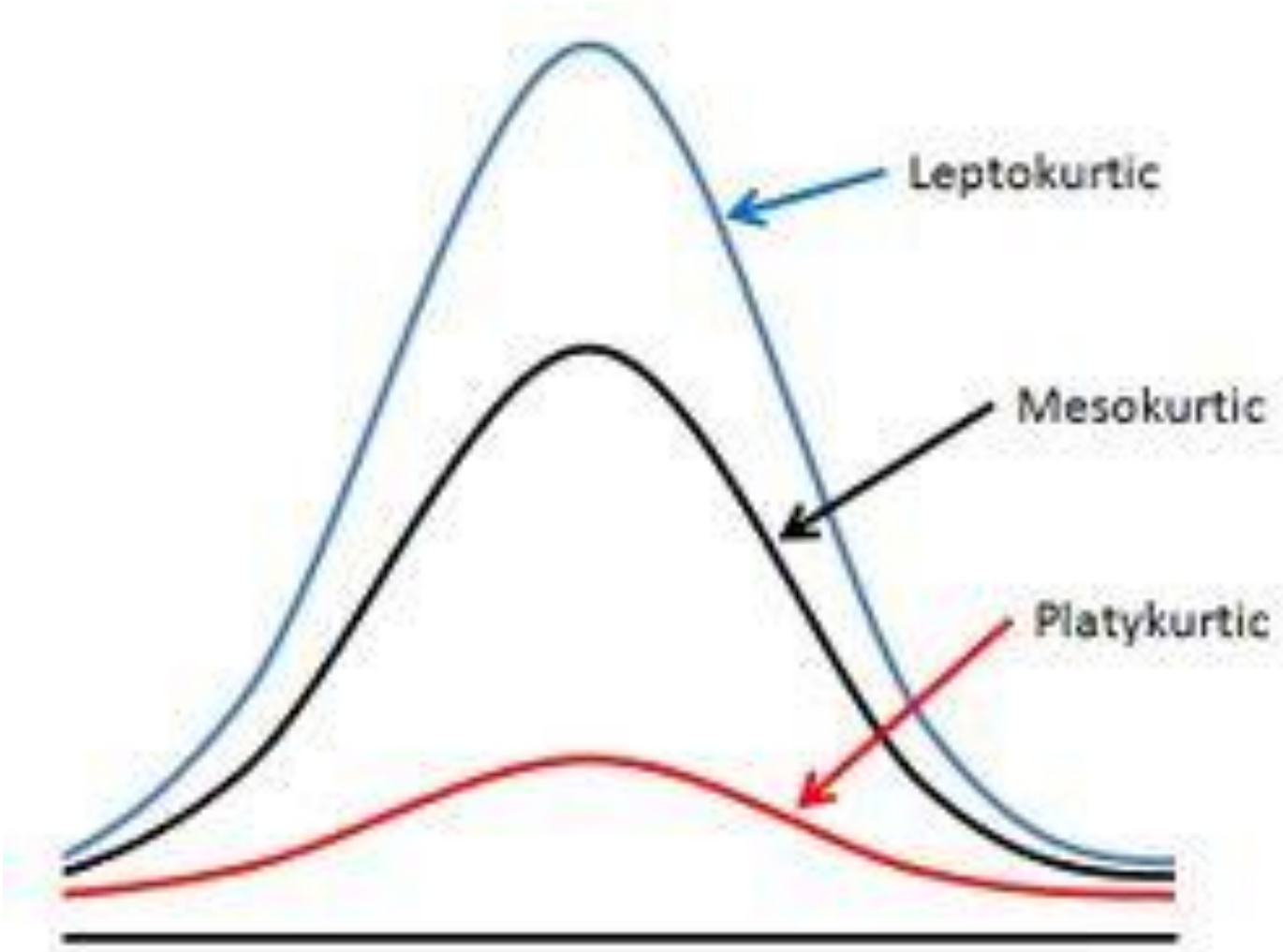
- **Standard Deviation**: The square root of the variance. It captures how far, on average, each score is from the mean value
  - We use this because it is more interpretable (However, we need the variance for many calculations)

What is skew?

- When extreme values “pull” the shape of the distribution

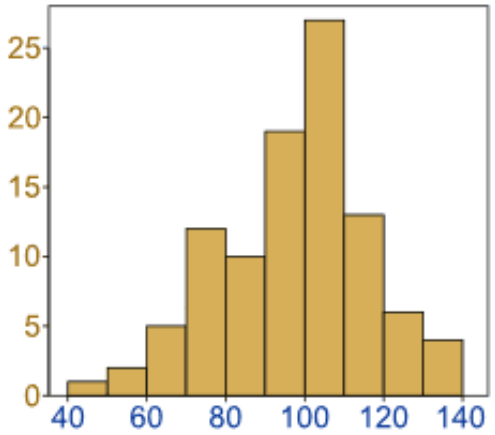


Kurtosis & interpretation

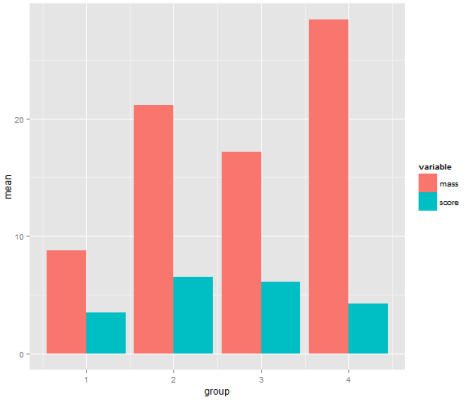


Visualizing descriptive statistics

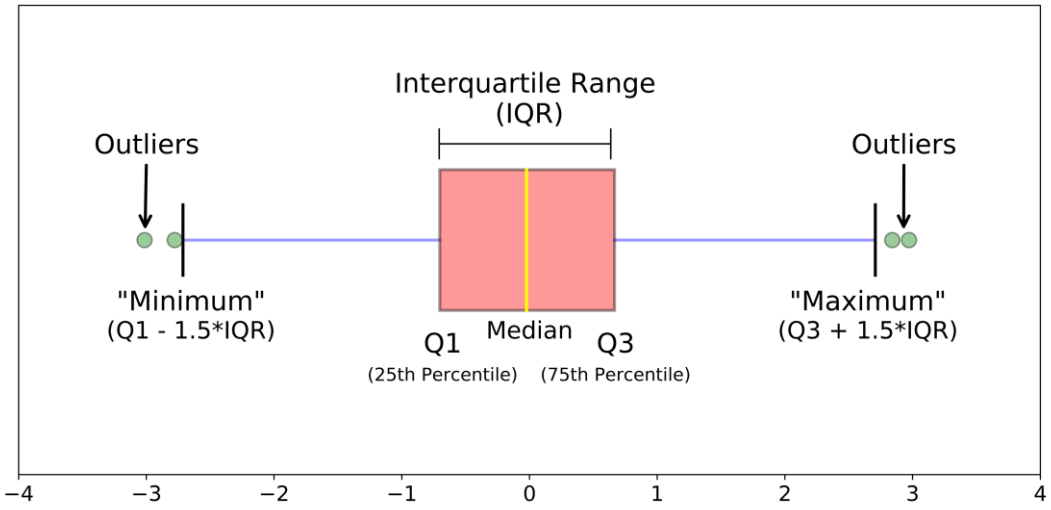
- Histograms
- Bar plots
- Boxplots



geom_histogram()



geom_bar()



geom_boxplot()

Part 3: Common terminology used in statistics

Common terms in statistics

- Degrees of freedom (df)
- Standard error
- Confidence intervals
- Test statistics (i.e. χ^2 test , t-test, F-test, etc.)

Degrees of freedom (df)

- The number of observations (or groups) that are free to vary in your sample.
  - For example: We have a of a sample of 3 with a sample mean of 10
    - In our sample we the following values: 5, 10, 15.
      - Because we know the mean, we only need 2 out of 3 of those values. Thus, 2 values are allowed to "vary" as much as possible – the " minus 1" is the only value we need to manipulate to get the mean value (or whatever your parameter of interest is).
- It conveniently makes our sample less variable than the population when we subtract that 1.

- Here's a great video explaining the concept (Starting at 7:08): https://youtu.be/nlm9gfso4mw?t=428

Standard error

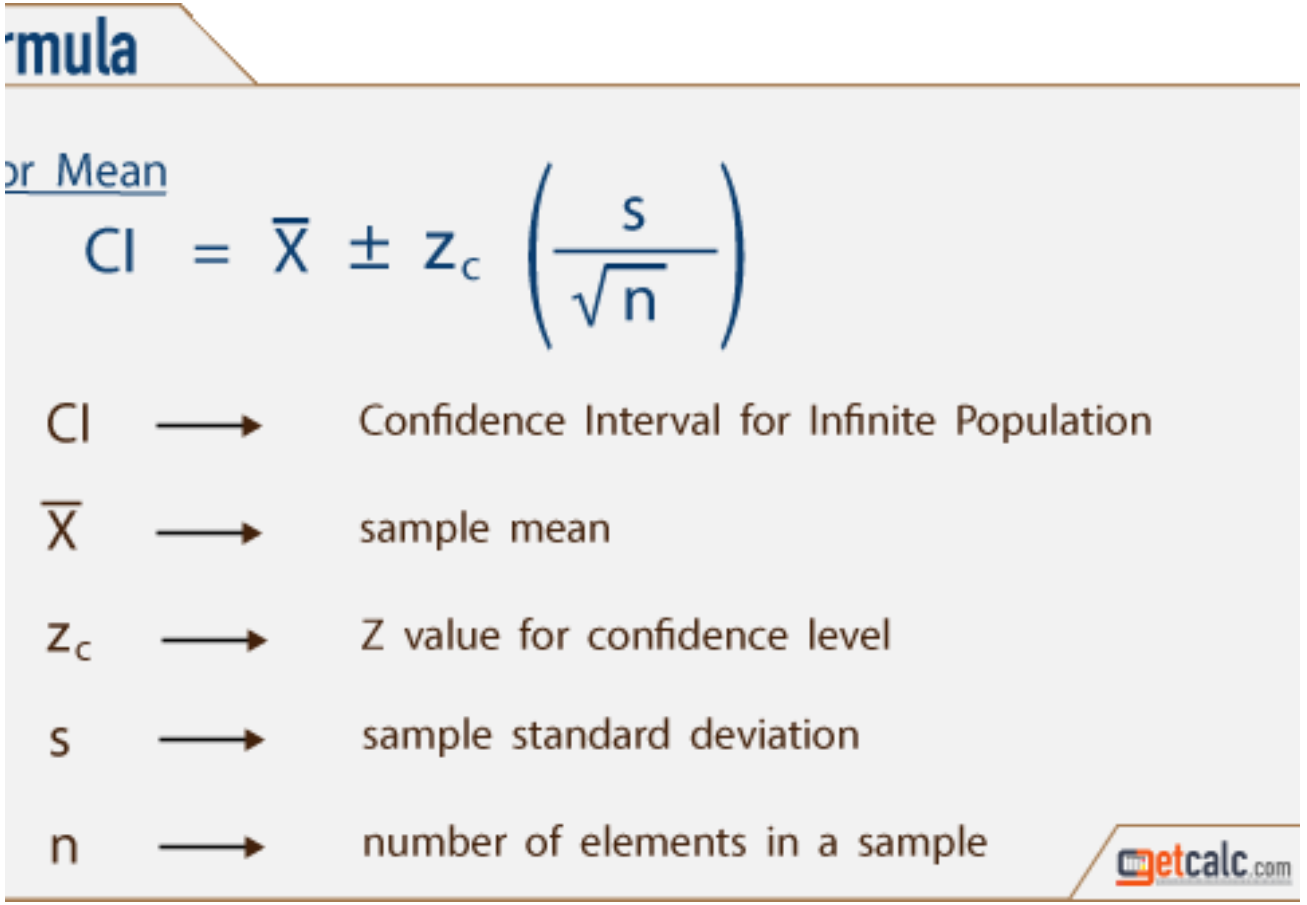
- Derived from the sampling distribution
  - Sampling distribution = If we got an infinite number of the same size samples of from a population, we would estimate the population parameter
- Used to understand the statistical accuracy of an estimate
- It is inferential
  - Inferential = making assumptions about the population from our sample
- If we continuously sample from the same population, the SD of the *sampling distribution* will be approximately equivalent to the SE.
- Simulation:
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html

$$SE = \frac{\sigma}{\sqrt{n}}$$

Confidence intervals

- Derived from the SE

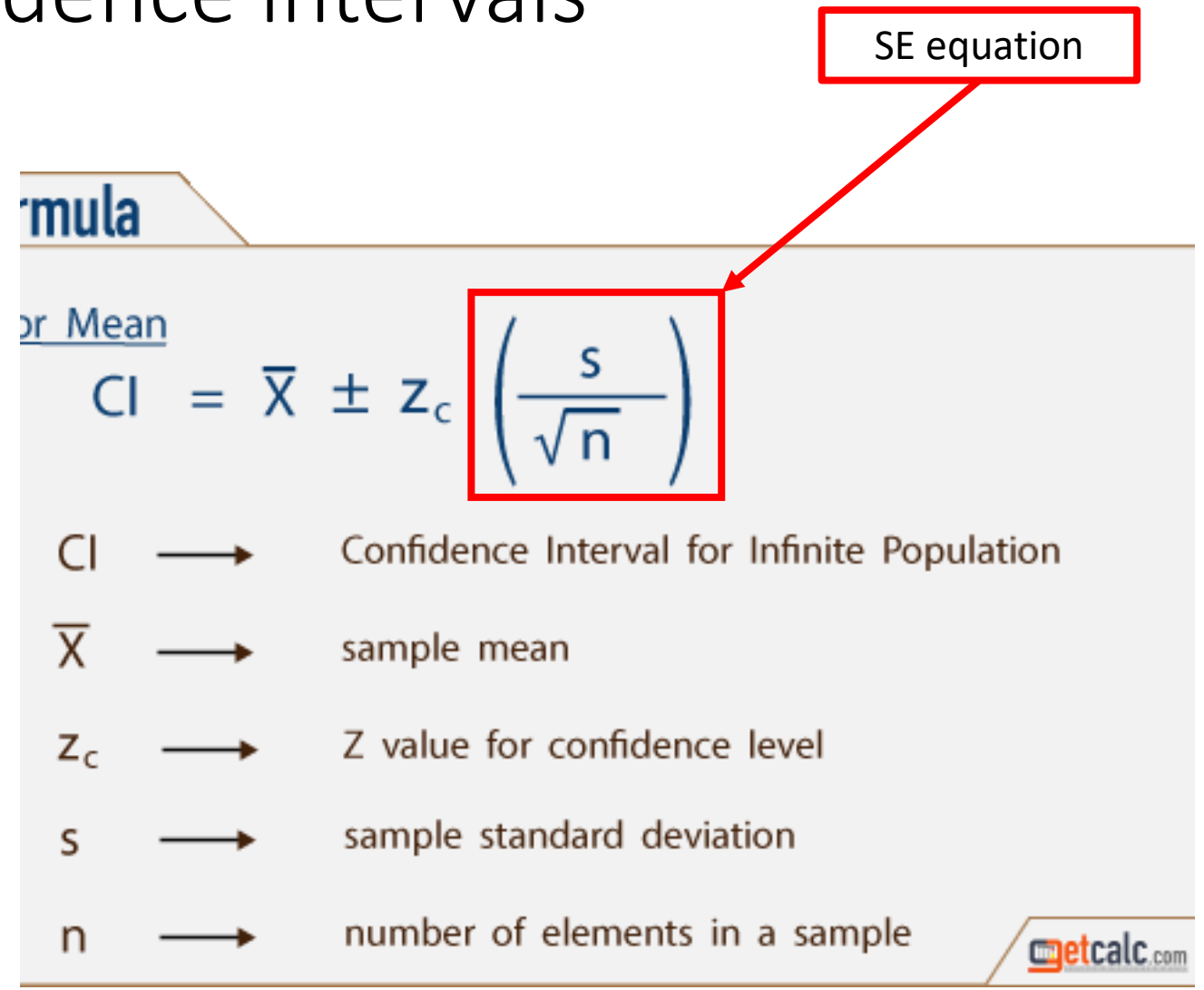
- A 95% Confidence interval means:
  If we ran this study 100 times, 95 of our confidence intervals will contain the parameter value

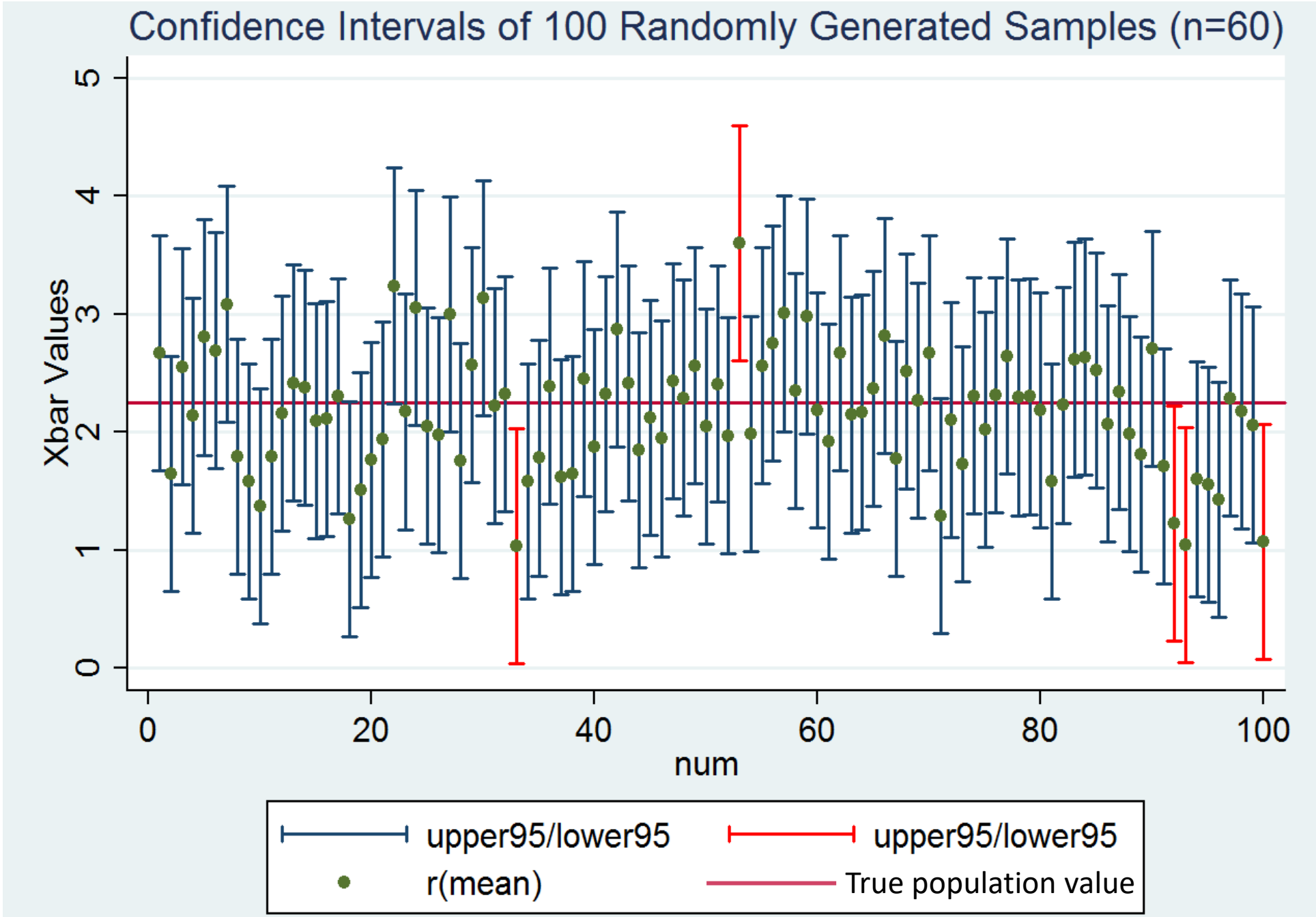


Confidence intervals

- Derived from the SE

- A 95% Confidence interval means:
  If we ran this study 100 times, 95 of our confidence intervals will contain the parameter value
  - **Pictured on the next slide**





* Notice how **5** of the 100 confidence intervals do NOT capture the true population value.

Test Statistics

- General formula:

$$\text{Test Statistic} = \frac{Explained}{Unexplained\ (Error)}$$

Test Statistics

- More specific formulas:

$$\text{t statistic} = \frac{Observed\ value - Null\ Value}{error}$$

OR

$$\text{F statistic} = \frac{Explained\ variation}{Unexplained\ variation\ (Error)}$$

Test Statistics

- More specific formulas:

This asks, "How different is our *point estimate* (i.e. mean) from zero?"

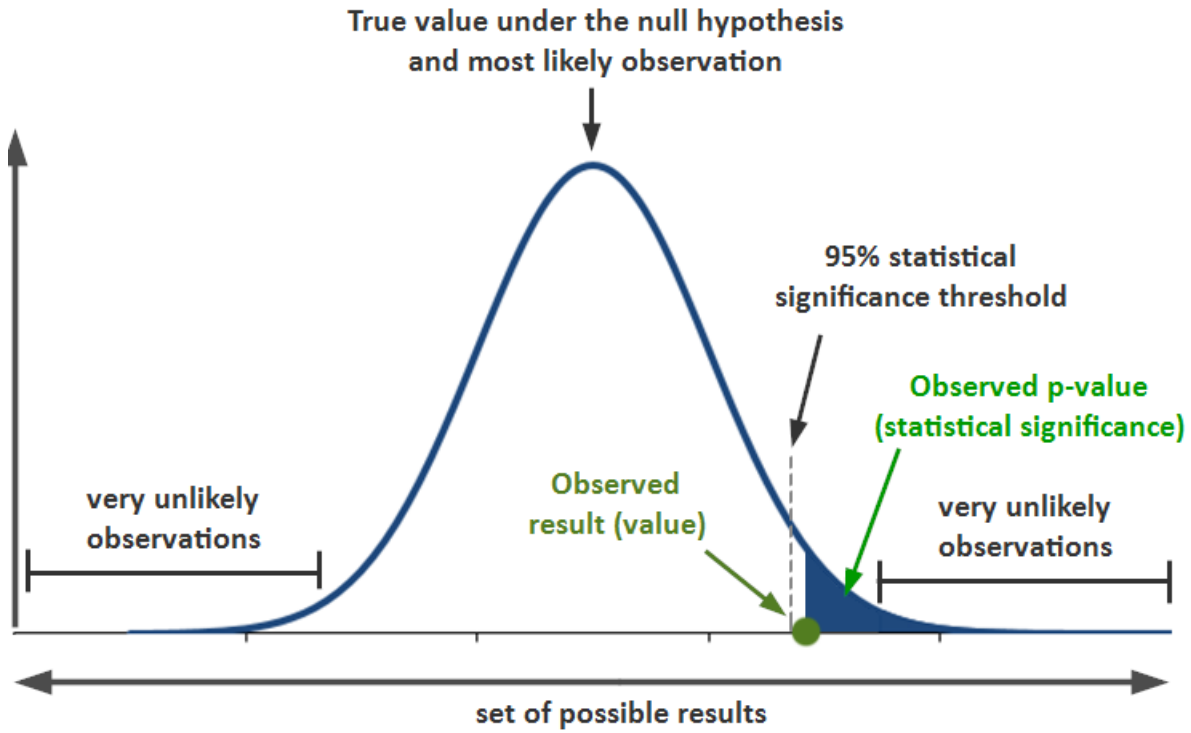
$$\text{t statistic} = \frac{Observed\ value - Null\ Value}{error}$$

OR

This asks, "How different from zero is the proportion of variance we are explaining (i.e. R^2)?"

$$\text{F statistic} = \frac{Explained\ variation}{Unexplained\ variation\ (Error)}$$

Test statistics and p-values



F-Distribution